What drives Renewable Energy Generation?

Cecilia Barradas

March 3rd, 2022

Introduction

Overview:

This project is the capstone for the course PH125.9x of Harvard X in order to earn the Data Science Certificate.

Goal:

This second part of the capstone consists in choosing a project of our own, finding a data set from the Kaggle or the UCI Machine Learning Repository, explore the data and apply machine learning techniques.

Data set:

Following the instructions, I browsed the publicly available data sets and chose a data set on crime in Vancouver. After exploring the data and creating data visualizations regarding crime types, times and places, I found that it had no machine learning modelling potential. I went back to choose another data set and started working with information on real estate sales in New York, but after a few explorations, I found it was terribly boring. After much browsing, I finally settled for a topic that I am deeply engaged with: Sustainable Energy. I had a data set obtained in another online course with data on energy generation, price, sales, pollution, and other demographic and political variables. After opening this data set, so many questions came to mind, which is a good indication of future work. I make the data set available in the Github website. It is important to notice that this data set does not provide technical documentation on the variables, therefore, we do not know if energy generation is measured in megawatts, gigawatts, or watt hours; if sales refer to electricity or monetary quantities, or if prices are in dollars or dollars per hour. The only certainties are that total salary in measured in US Dollars and government incentives in number. As a result, this data set is good to train machine learning models and see the importance of the predictors on the dependent variable, but it has very low real life interpretability.

Key Steps:

1. Loaded the data set
2. Explore the data for relationships between the variables to choose the best ones for machine learning modelling.
3. Divided the data into training and testing sets, 80% of the data went into the train set and 20% to the test set.

4. As the dependent variable was continuous, 3 machine learning techniques were chosen to train the data: Linear Regression, Decision Trees and Random Forest. Each was trained with 3 models containing: all variables, 9 variables (a combination of significant variables and theoretically important variables), and the 6 most significant variables.
5. A table comparing the RMSE of all the models and techniques is shown in the results to show the best method, along with an explanation on the most important variables in each model.

Method

The following libraries were used:tidyverse, caret, data.table, ggplot2, dplyr, knitr, RColorBrewer, rmarkdown, pdftools, kableExtra, rpart, rpart.plot, randomForest, lubridate, purrr, e1071, scales, ggpubr, ggrepel, parameters and insight.

1.Data Preparation

variable	class	first__values
STATE	character	AK, AK, AK, AK, AK, AK
YEAR	integer	2000, 2001, 2002, 2003, 2004, 2005
GenTotal	numeric	9.807319179, 10.64461169, 10.54143158, 9.786962825, 9.899389509, 9.858373531
GenHydro	numeric	0.162724751, 0.199542066, 0.212691278, 0.249661289, 0.229521213, 0.22259661
GenSolar	numeric	0, 0, 0, 0, 0, 0
GenTotalRenewable	numeric	0.162724751, 0.199682937, 0.214335059, 0.250453087, 0.230873807, 0.223427123
GenSolarBinary	integer	0, 0, 0, 0, 0, 0
GenTotalRenewableBinary	integer	1, 1, 1, 0, 0, 0
AllSourcesCO2	numeric	7.260626876, 7.311766619, 6.93834174, 6.153593414, 7.246256285, 6.541565609
AllSourcesSO2	numeric	0.02239593, 0.011925094, 0.0114771, 0.006750341, 0.00659179, 0.006048441
AllSourcesNOx	numeric	0.028911283, 0.027725251, 0.029431098, 0.024219087, 0.037735191, 0.025177106
EPriceResidential	numeric	11.45, 12.12, 12.05, 11.98, 12.44, 13.3
EPriceCommercial	numeric	9.77, 10.29, 10.13, 10.49, 10.99, 11.56
EPriceIndustrial	numeric	7.56, 7.61, 7.65, 7.86, 8.33, 9.29
EPriceTransportation	numeric	0, 0, 0, 0, 0, 0
EPriceTotal	numeric	10.08, 10.54, 10.46, 10.5, 10.99, 11.72
EsalesResidential	numeric	0.349336814, 0.346798727, 0.353530489, 0.357139211, 0.356208119, 0.348689597
EsalesCommercial	numeric	0.421184489, 0.41968783, 0.409521637, 0.444494851, 0.449272728, 0.455793427
EsalesIndustrial	numeric	0.19522276, 0.197911105, 0.199040745, 0.198365938, 0.194519152, 0.195516976
EsalesTransportation	numeric	0, 0, 0, 0, 0, 0
EsalesOther	numeric	0.034255937, 0.035602338, 0.03790713, 0, 0, NA
EsalesTotal	numeric	8.458760522, 8.608923222, 8.513567528, 8.590290441, 8.779675567, 8.862909488
CumlFinancial	integer	1, 1, 1, 1, 1, 2
CumlRegulatory	integer	1, 1, 1, 1, 1, 1
Total.salary	numeric	17.64608569, 18.69197034, 19.64402117, 20.39238357, 21.44678279, 22.71281629
presidential.results	integer	0, 0, 0, 0, 0, 0
Import	integer	0, 0, 0, 0, 0, 0

The data set has 699 observations and 27 variables. At first glance, removing NA's seemed adequate, yet removing all NAs leaves a data set with only 249 observations.

Therefore, removing variables with many NAs like GenSolarBinary and GenTotalRenewableBinary that will not be useful for this exercise, seemed more appropriate. Also, we will remove AllSourcesCO2, AllSourcesSO2 and AllSourcesNOx because they contain 50 NA's each, and even though very interesting questions can be answered with them, they will be matter of a later exercise. Finally, we will remove, ESalesOther, as it has 449 NA's and it wont bring any insight; and Import, which leaves a final data set of 699 observations and 20 variables.

variable	class	first__values
STATE	character	AK, AK, AK, AK, AK, AK
YEAR	integer	2000, 2001, 2002, 2003, 2004, 2005
GenTotal	numeric	9.807319179, 10.64461169, 10.54143158, 9.786962825, 9.899389509, 9.858373531
GenHydro	numeric	0.162724751, 0.199542066, 0.212691278, 0.249661289, 0.229521213, 0.22259661
GenSolar	numeric	0, 0, 0, 0, 0, 0
GenTotalRenewable	numeric	0.162724751, 0.199682937, 0.214335059, 0.250453087, 0.230873807, 0.223427123
EPriceResidential	numeric	11.45, 12.12, 12.05, 11.98, 12.44, 13.3
EPriceCommercial	numeric	9.77, 10.29, 10.13, 10.49, 10.99, 11.56
EPriceIndustrial	numeric	7.56, 7.61, 7.65, 7.86, 8.33, 9.29
EPriceTransportation	numeric	0, 0, 0, 0, 0, 0
EPriceTotal	numeric	10.08, 10.54, 10.46, 10.5, 10.99, 11.72
EsalesResidential	numeric	0.349336814, 0.346798727, 0.353530489, 0.357139211, 0.356208119, 0.348689597
EsalesCommercial	numeric	0.421184489, 0.41968783, 0.409521637, 0.444494851, 0.449272728, 0.455793427
EsalesIndustrial	numeric	0.19522276, 0.197911105, 0.199040745, 0.198365938, 0.194519152, 0.195516976
EsalesTransportation	numeric	0, 0, 0, 0, 0, 0
EsalesTotal	numeric	8.458760522, 8.608923222, 8.513567528, 8.590290441, 8.779675567, 8.862909488
CumlFinancial	integer	1, 1, 1, 1, 1, 2
CumlRegulatory	integer	1, 1, 1, 1, 1, 1
Total.salary	numeric	17.64608569, 18.69197034, 19.64402117, 20.39238357, 21.44678279, 22.71281629
presidential.results	integer	0, 0, 0, 0, 0, 0

Energy is such a vital sector of our lives. We can not live without energy. This data set talks about energy generation, price, sales and the variables that have inference in them, like governmental incentives, salary level in the population and politics. At a first glance many questions arise: What is the relationship between government incentives and renewable energy generation?, between generation and price?, between price and sales?, between salary and sales and production? Have energy prices risen or decreased with time, with renewable energy production? Will they rise or decrease in the future?

In order to answer some of these questions, this data set needs to be averaged in order to have one averaged point per year and make clean visualizations, if not we will have 50 observations per year(one per State) clouding clear results. We create a data frame with this new averaged variables, removing the YEAR repetition.

variable	class	first__values
STATE	character	AK, AK, AK, AK, AK, AK
YEAR	integer	2000, 2001, 2002, 2003, 2004, 2005
GenTotal	numeric	9.807319179, 10.64461169, 10.54143158, 9.786962825, 9.899389509, 9.858373531
GenHydro	numeric	0.162724751, 0.199542066, 0.212691278, 0.249661289, 0.229521213, 0.22259661
GenSolar	numeric	0, 0, 0, 0, 0, 0
GenTotalRenewable	numeric	0.162724751, 0.199682937, 0.214335059, 0.250453087, 0.230873807, 0.223427123
EPriceResidential	numeric	11.45, 12.12, 12.05, 11.98, 12.44, 13.3
EPriceCommercial	numeric	9.77, 10.29, 10.13, 10.49, 10.99, 11.56
EPriceIndustrial	numeric	7.56, 7.61, 7.65, 7.86, 8.33, 9.29
EPriceTransportation	numeric	0, 0, 0, 0, 0, 0
EPriceTotal	numeric	10.08, 10.54, 10.46, 10.5, 10.99, 11.72
EsalesResidential	numeric	0.349336814, 0.346798727, 0.353530489, 0.357139211, 0.356208119, 0.348689597
EsalesCommercial	numeric	0.421184489, 0.41968783, 0.409521637, 0.444494851, 0.449272728, 0.455793427
EsalesIndustrial	numeric	0.19522276, 0.197911105, 0.199040745, 0.198365938, 0.194519152, 0.195516976
EsalesTransportation	numeric	0, 0, 0, 0, 0, 0
EsalesTotal	numeric	8.458760522, 8.608923222, 8.513567528, 8.590290441, 8.779675567, 8.862909488
CumlFinancial	integer	1, 1, 1, 1, 1, 2
CumlRegulatory	integer	1, 1, 1, 1, 1, 1
Total.salary	numeric	17.64608569, 18.69197034, 19.64402117, 20.39238357, 21.44678279, 22.71281629
presidential.results	integer	0, 0, 0, 0, 0, 0

This leaves a data set of 16 variables with 14 observations, one per year from 2000 to 2013.

2. Data Exploration, Visualization and Insights

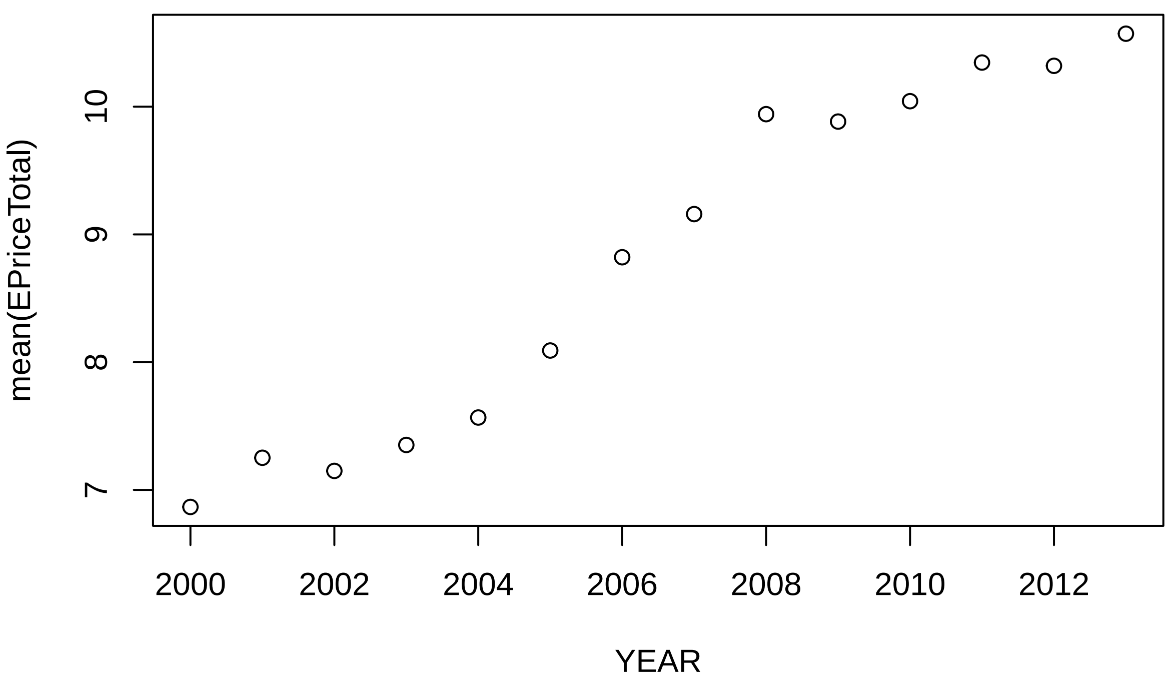
Theory would follow that an increase in government incentives would provoke a surge in energy generation that would be accompanied by a decrease in prices due to increased demand. This in turn will increase consumption due to more affordable energy prices. Another path is that increase in income causes increase in consumption. We will try to see with this data set if the theory can be proven or debunked.

An Increase in Government Incentives -> Increase in Energy Generation-> Decrease in Prices(more offer)-> Increase in Consumption

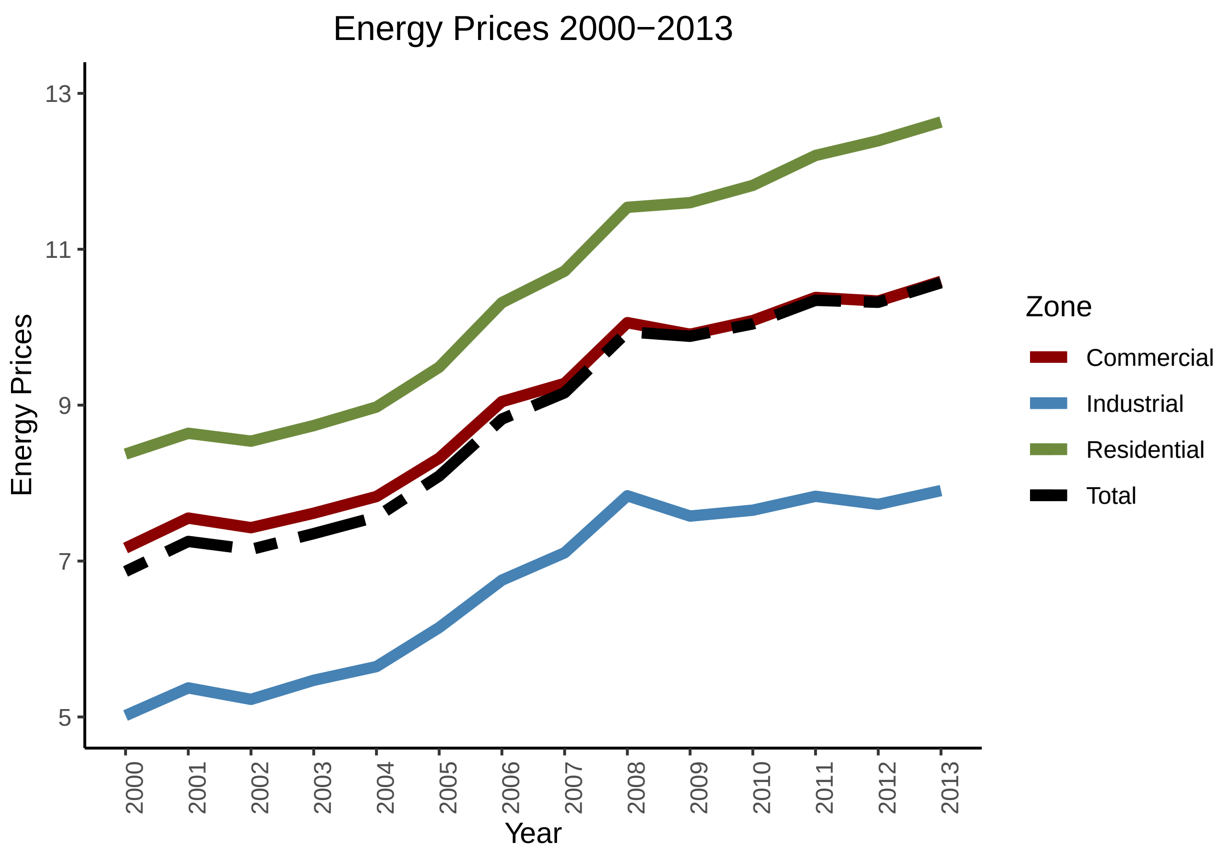
Increase in income -> Increase in Energy Consumption

2.1. Price

With this averaged data set, we can start making exploratory visualizations. First quick question is: Has the price of energy increased or decreased over the years. Plotting the average price per year quickly shows that energy prices have increased from 2000 to 2013.



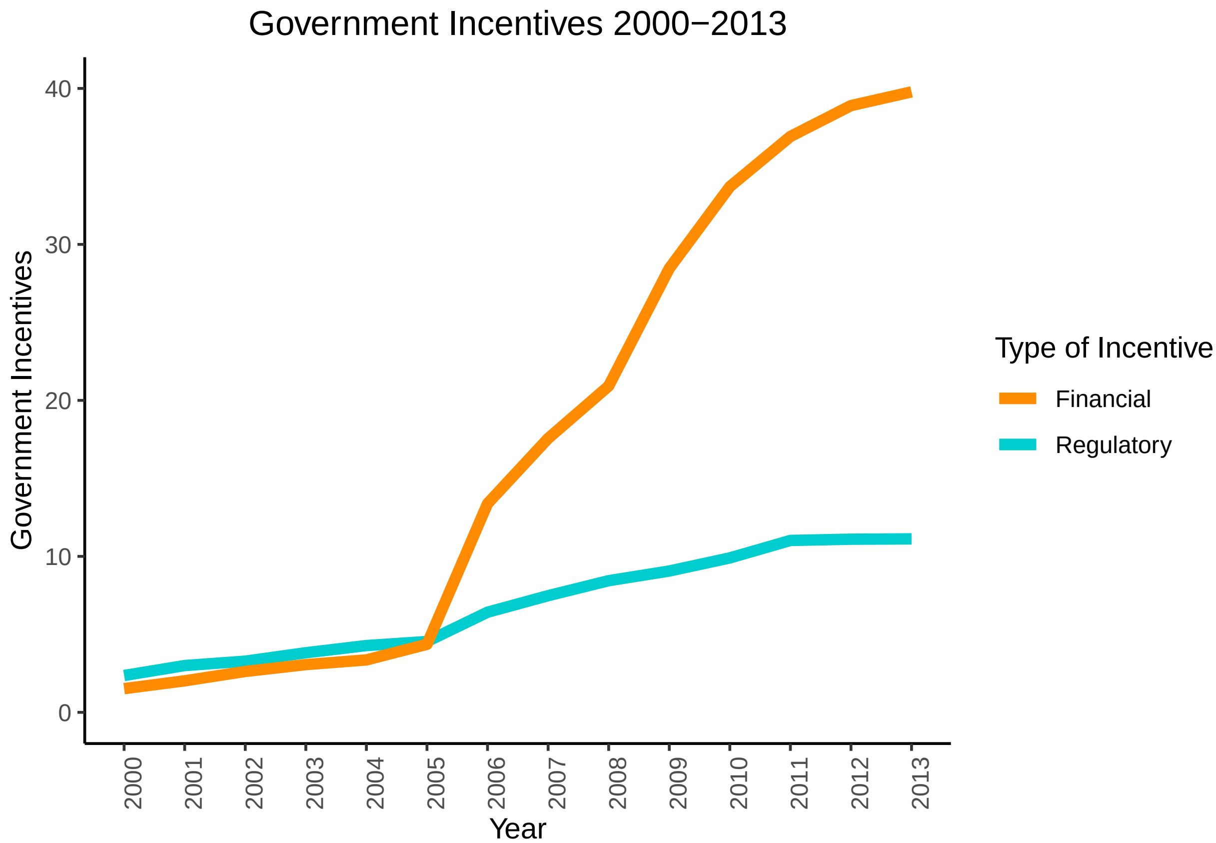
To check that this trend is followed by all human activities, we plot energy prices in residential, commercial and industrial zones.



We can clearly see that there has been a price upward trend in the time studied.

2.2 Government Incentives

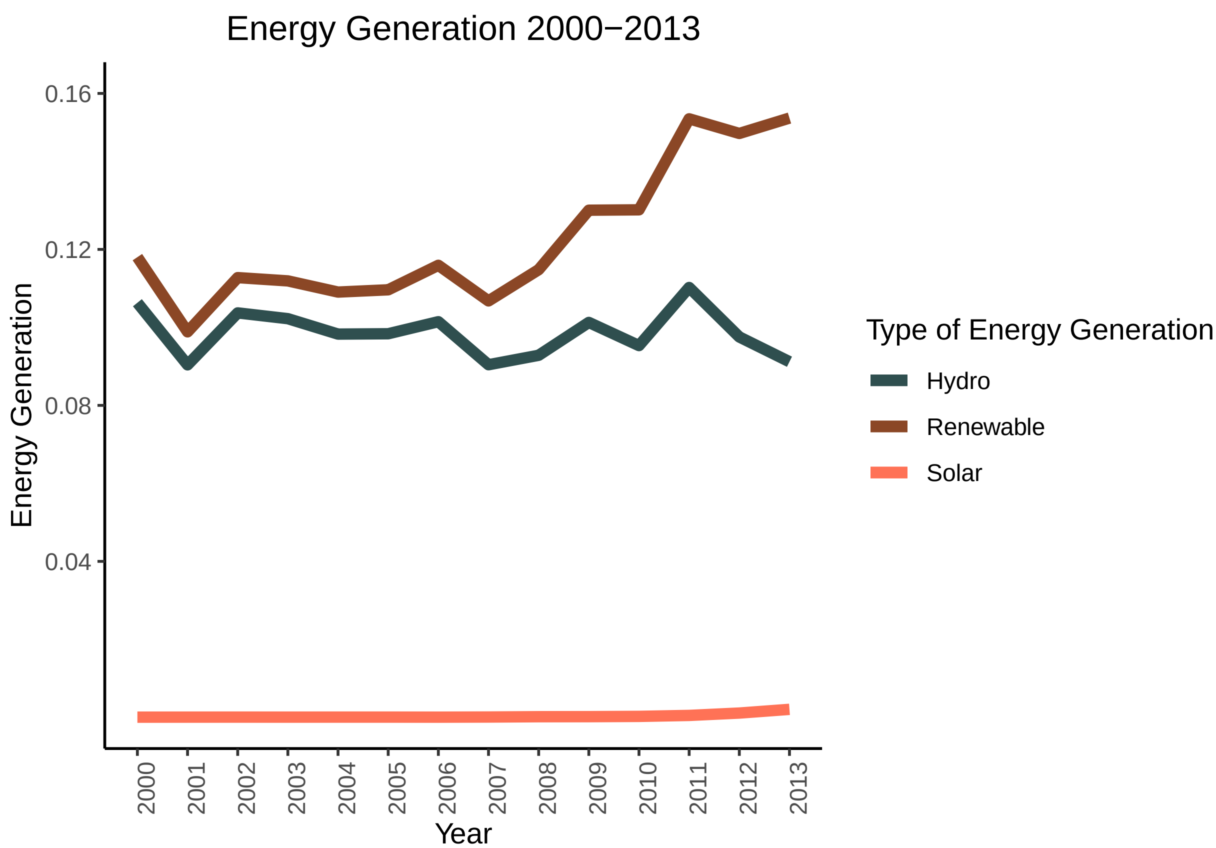
Another upward trend, specially after 2005 was the implementation of government incentives for energy generation.



This is consistent historically with the policies implemented in 2009, when President Obama and Congress worked together to combat a severe economic recession by passing a massive economic stimulus plan. Among its many provisions, the American Recovery and Reinvestment Act of 2009 provided US$90 billion to promote clean energy. The bill's clean energy package was dubbed the "biggest energy bill in history".

2.3 Energy Generation

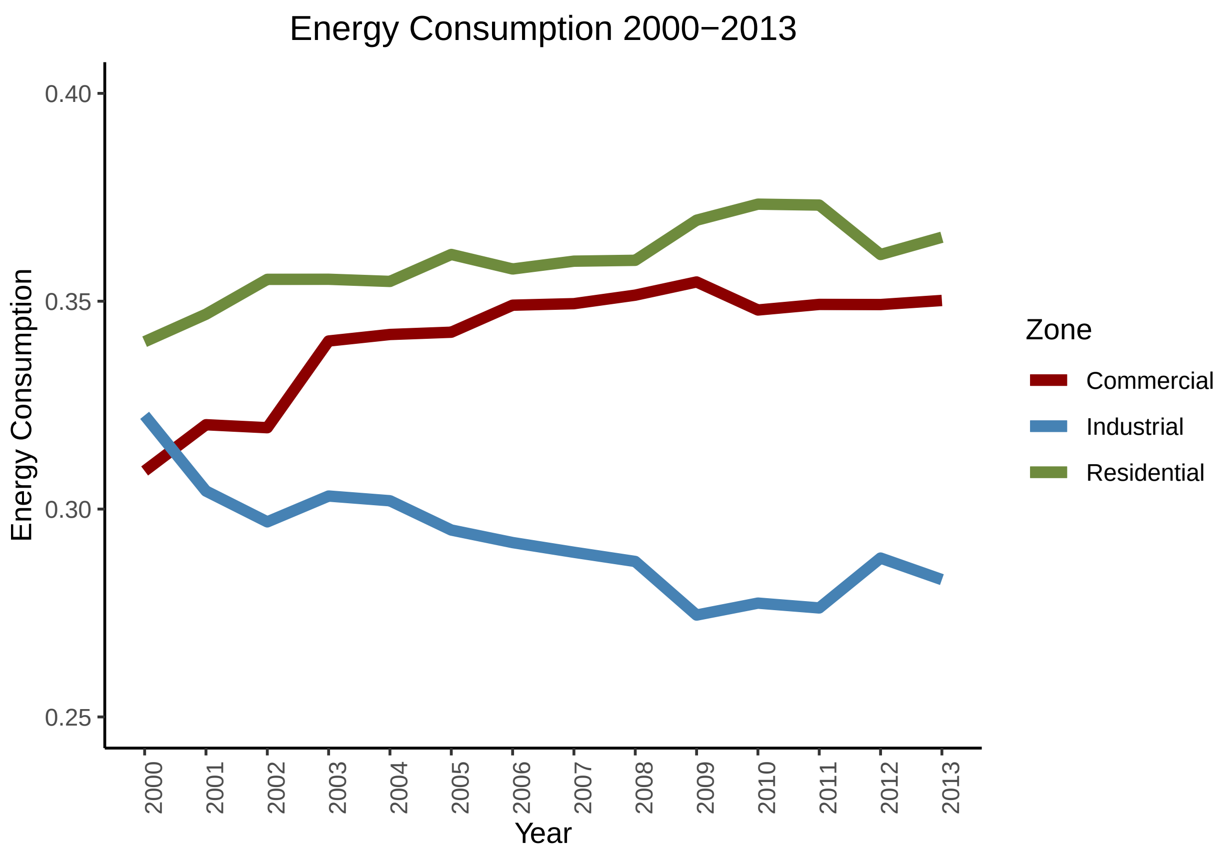
Was this followed by higher Energy generation? In terms of total energy, we do not see an increase in generation that would follow the increase in government incentives. On the contrary, from 2010, there seems to be a downward trend. To dig deeper, we will see the subsections of energy generation: Solar, Hydro and Renewable (taking into account that both hydro and solar are renewable energy generation sources).



The relationship is not as straightforward, we can see that there is an upward trend in total renewable Energy generation, which must be caused by an external factor such as wind power, which is not included in this data set, because Hydroelectric power and solar do not seem to follow the upward trend.

2.4 Energy Consumption

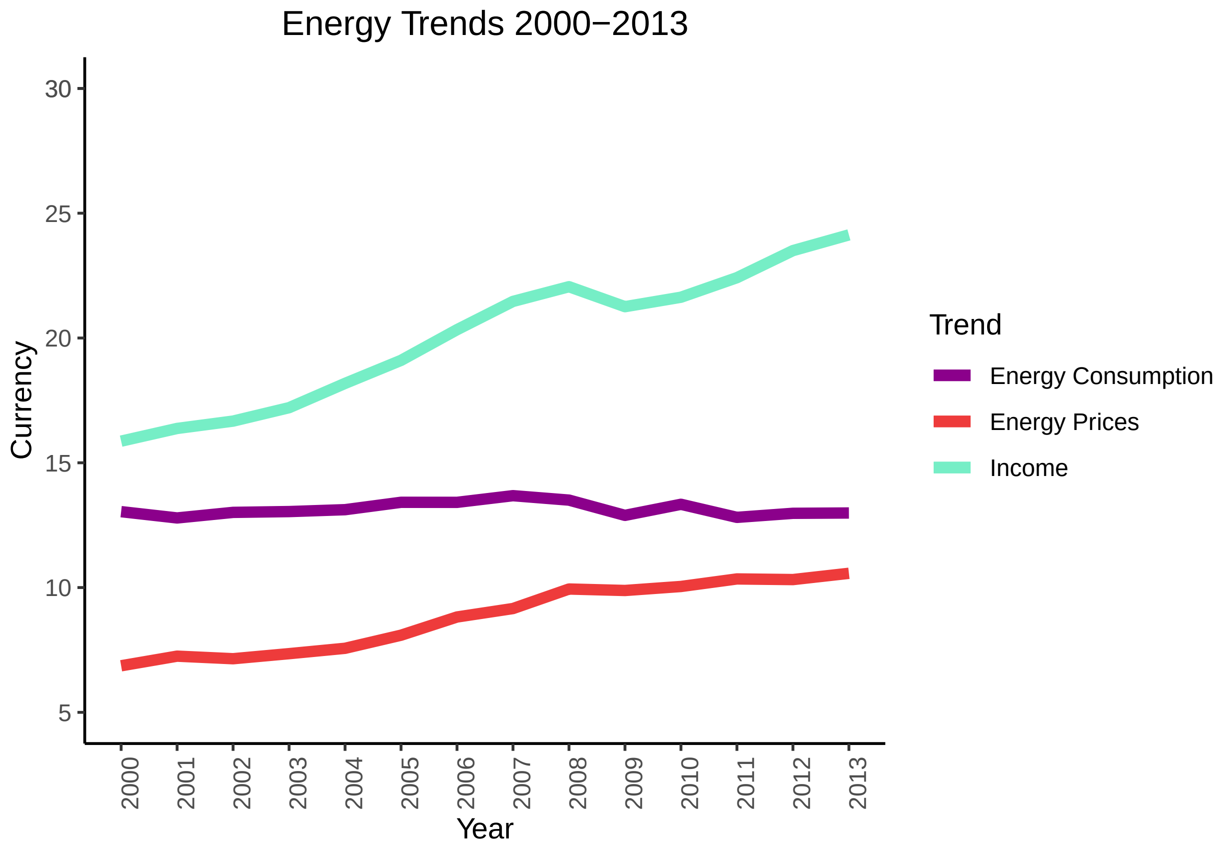
Since there was not more energy generated, was there more being consumed? In this case we will use energy sales as a proxy for energy consumption.



We can see a slight increase in the consumption of energy for the residential and commercial areas and a decrease in the industrial zones. This can be explained as the initial steps of the industry to be energy independent and generate its own power.

2.5 Salary

Finally, we will investigate if population income has an impact in energy prices.



This last graph shows that Energy Consumption did not variate much in the 13 years of the study, it was pulled up by income but pulled down by rising energy prices.

3. Data Modeling

For the machine learning models, we will use the original data set with the 699 observations. First we will divide the data in train and test set. After many errors, I realized there was one NA in Epricetransportation that kept interfering, therefore I dropped the NA's, leaving a data set of 698 observations, 558 in the training set and 140 in the test set.

3.1 Linear Regression

To know which variables to include in the model, we measure their importance with a simple linear regression + elimination. A first model includes all variables except State and YEAR

Parameter	Coefficient	SE	95% CI	t(680)	p
(Intercept)	1.90e-03	0.08	(-0.15, 0.15)	0.02	0.981
GenTotal	-1.47e-05	1.61e-04	(-3.31e-04, 3.02e-04)	-0.09	0.927
GenHydro	1.02	7.09e-03	(1.01, 1.04)	144.44	< .001
GenSolar	2.19	0.77	(0.68, 3.69)	2.86	0.004
EPriceResidential	2.69e-03	6.08e-03	(-9.24e-03, 0.01)	0.44	0.658
EPriceCommercial	-3.97e-03	5.84e-03	(-0.02, 7.51e-03)	-0.68	0.498
EPriceIndustrial	5.60e-04	3.88e-03	(-7.06e-03, 8.18e-03)	0.14	0.885
EPriceTransportation	-1.29e-03	3.46e-04	(-1.97e-03, -6.08e-04)	-3.72	< .001

Parameter	Coefficient	SE	95% CI	t(680)	p
EPriceTotal	1.37e-03	0.01	(-0.03, 0.03)	0.09	0.927
EsalesResidential	-0.13	0.09	(-0.30, 0.04)	-1.50	0.135
EsalesCommercial	0.03	0.08	(-0.13, 0.20)	0.38	0.701
EsalesIndustrial	0.03	0.09	(-0.14, 0.20)	0.34	0.736
EsalesTransportation	-1.34	0.69	(-2.70, 0.02)	-1.94	0.053
EsalesTotal	-5.45e-05	7.87e-04	(-1.60e-03, 1.49e-03)	-0.07	0.945
CumlFinancial	5.79e-04	8.37e-05	(4.15e-04, 7.44e-04)	6.92	< .001
CumlRegulatory	2.79e-04	4.28e-04	(-5.62e-04, 1.12e-03)	0.65	0.515
Total salary	1.67e-03	4.72e-04	(7.45e-04, 2.60e-03)	3.54	< .001
presidential results	5.60e-04	3.34e-03	(-6.00e-03, 7.12e-03)	0.17	0.867

This first model has a very high adjusted R2 of .9741 and shows that the most significant variables are GenHydro, EPriceTransportation, CumlFinancial and Total.Salary; and to a lesser extent GenSolar and EsalesTransportation. These outcome does not fully match conventional theory or the information observed in the data exploration, but allows for the creation of a model with the significant variables. Reducing the model to include less variables without losing significance in the adjusted r-squared, gives a model with 9 variables (the most significant according to the linear regression and theoretical hypothesis.)

Parameter	Coefficient	SE	95% CI	t(688)	p
(Intercept)	-0.03	0.01	(-0.05, -7.95e-03)	-2.75	0.006
GenTotal	3.10e-04	1.44e-04	(2.73e-05, 5.93e-04)	2.15	0.032
GenHydro	1.02	6.92e-03	(1.01, 1.03)	147.22	< .001
GenSolar	2.54	0.74	(1.07, 4.00)	3.40	< .001
EPriceTransportation	-1.67e-03	3.35e-04	(-2.33e-03, -1.02e-03)	-4.99	< .001
EPriceTotal	6.78e-04	5.16e-04	(-3.35e-04, 1.69e-03)	1.31	0.189
EsalesTransportation	-1.34	0.49	(-2.30, -0.38)	-2.73	0.006
EsalesTotal	-5.87e-04	5.58e-04	(-1.68e-03, 5.08e-04)	-1.05	0.293
CumlFinancial	6.09e-04	6.06e-05	(4.90e-04, 7.28e-04)	10.06	< .001
Total salary	2.16e-03	3.88e-04	(1.39e-03, 2.92e-03)	5.55	< .001

Interestingly the importance of transportation continues to be significant, while the totals are not, therefore, we run a third model, with only the 6 significant variables without losing value in the adjusted r-squared.

Parameter	Coefficient	SE	95% CI	t(691)	p
(Intercept)	-0.03	6.50e-03	(-0.04, -0.02)	-4.35	< .001
GenHydro	1.02	6.78e-03	(1.00, 1.03)	149.82	< .001
GenSolar	2.61	0.74	(1.15, 4.06)	3.51	< .001
EPriceTransportation	-1.89e-03	3.23e-04	(-2.52e-03, -1.26e-03)	-5.85	< .001
EsalesTransportation	-1.08	0.47	(-2.01, -0.15)	-2.29	0.023
CumlFinancial	6.06e-04	6.07e-05	(4.87e-04, 7.25e-04)	9.98	< .001
Total salary	2.40e-03	3.47e-04	(1.72e-03, 3.08e-03)	6.92	< .001

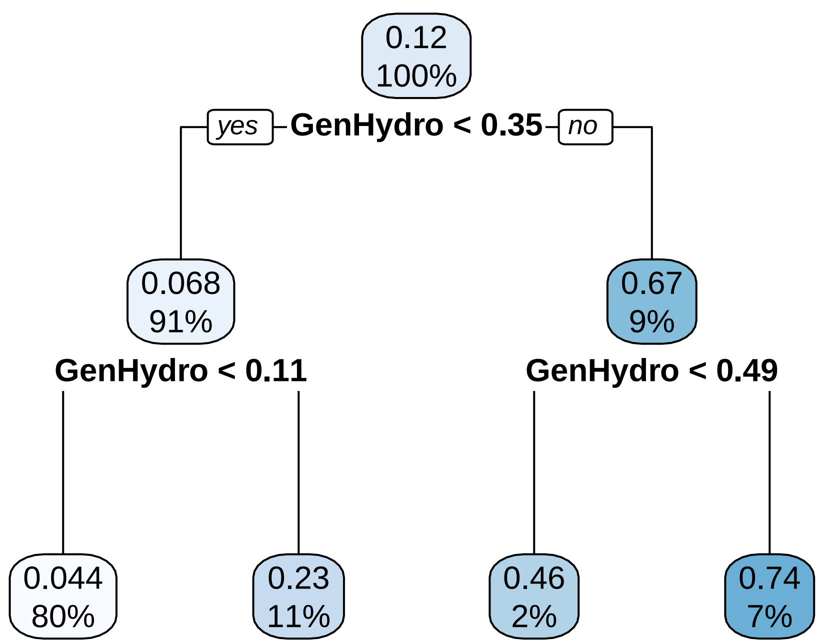
This last model shows all significant variables that impact the dependent variable with the right sign. We will now test their predictive power. We run the 3 models on the train set and test them on the test set with the goal of obtaining the lowest RMSE. RMSE (Root Mean Square Error) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit. The lowest the RMSE the better the model is. Since we are doing regression and not categorization, we can not measure accuracy.

Model	RMSE
Linear Regression All Variables	0.0005619
Linear Regression 9 Variables	0.0015678
Linear Regression 6 Variables	0.0021127

The three models have similar RMSE's, the reason behind the first being the lowest is the number of predictors. Adding extra predictors can improve RMSE substantially, but it wont improve the model when the added predictors are highly correlated to each other. Therefore we will again keep the third model (the one with the 6 significant variables) as the best model.

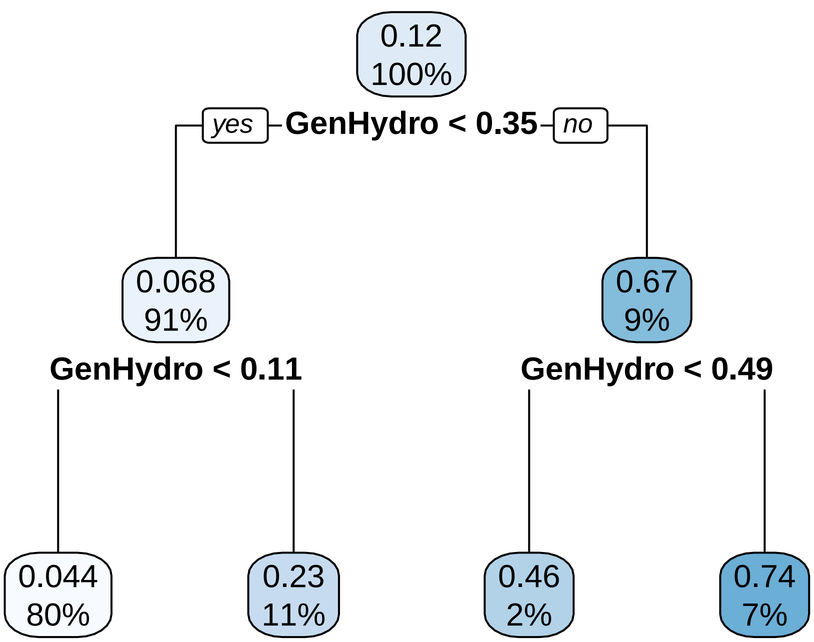
3.2 Regression Trees

We try visualizing the regression trees in its entirety in order to have an idea of the importance of variables. Again, a first tree including all variables, showed Gen Hydro as the most important variable, up to the point of being the only variable in the tree.



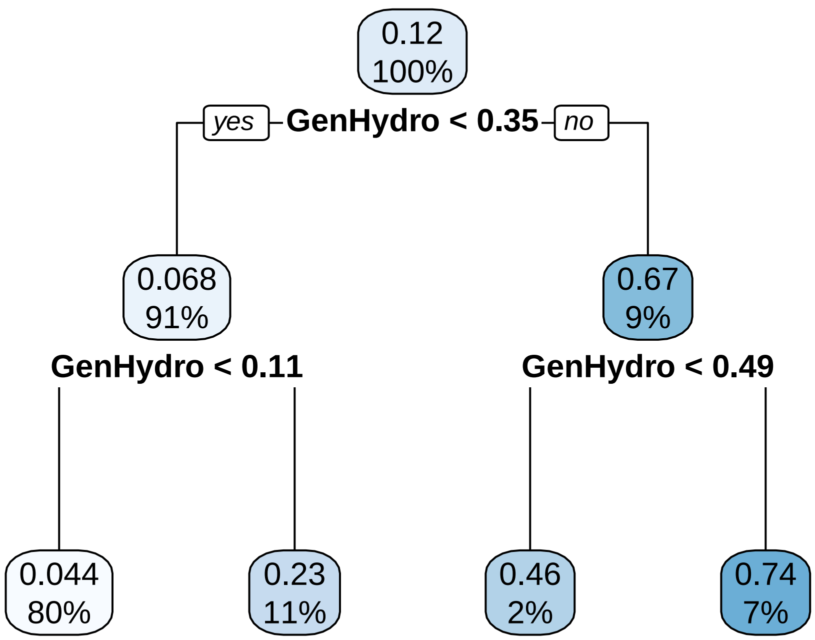
	Overall
CumlFinancial	0.0901541
CumlRegulatory	0.1684056
EPriceCommercial	0.2787946
EPriceResidential	0.2617276
EPriceTotal	0.1569509
EsalesCommercial	0.2068412
EsalesIndustrial	0.1674651
EsalesTotal	0.2116634
GenHydro	2.1476160
GenTotal	0.4809290
GenSolar	0.0000000
EPriceIndustrial	0.0000000
EPriceTransportation	0.0000000
EsalesResidential	0.0000000
EsalesTransportation	0.0000000
Total.salary	0.0000000

We try with the second model of 9 variables, but the result is identical. GenHydro is still the only variable shown in the regression tree.



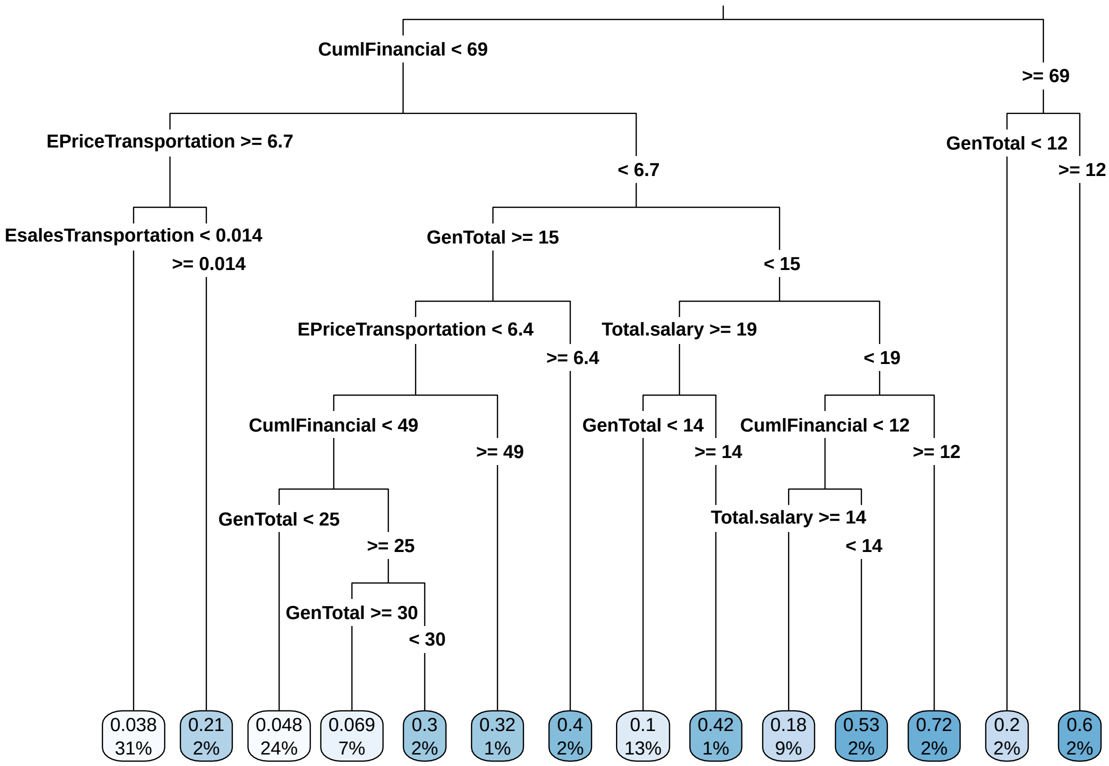
	Overall
CumlFinancial	0.2961419
EPriceTotal	0.2761356
EPriceTransportation	0.0280144
EsalesTotal	0.3718684
GenHydro	2.1476160
GenSolar	0.0996228
GenTotal	0.4809290
EsalesTransportation	0.0000000
Total.salary	0.0000000

The last model with only the 6 significant variables according to the linear regression model gives the same exact result.



	Overall
CumlFinancial	0.2961419
EPriceTransportation	0.0994815
EsalesTransportation	0.1259654
GenHydro	2.1476160
GenSolar	0.1024443
Total.salary	0.1139959

To try to see beyond this variable, we remove it from the model and we include GenTotal. As GenHydro is a source of Renewable Energy, there is a high correlation between these two variables that is overrunning this regression tree. Removing it from the model makes it more balanced and it may allow for other variables to show.



	Overall
CumlFinancial	1.8398162
EPriceTransportation	0.9598885
EsalesTransportation	0.9958033
GenSolar	0.0882215
GenTotal	2.1987321
Total.salary	1.0663450

We train the 4 models to see their predictive power.

Model	RMSE
Regression Tree All Variables	0.0674774
Regression Tree 9 Variables	0.0674774
Regression Tree 6 Variables	0.0674774
Regression Tree No GenHydro	0.1590990

RMSE's in the first 3 models are identical. RMSE increases heavily in the fourth model because we are removing a variable that the model considers most important. Here, We can experience the trade off between visibility and representativity. GenHydro is such an important variable that clouds others in a regression tree and removing it from the model makes it loose predictive power.

3.3 Random Forest

The last machine learning method we will use is random forest, as with the previous two models, we will try with the 4 models, all variables, 9 variables, 6 significant variables and GenHydro excluded.

Model	RMSE
Random Forest All Variables	0.0232776
Random Forest 9 Variables	0.0011295
Random Forest 6 Variables	0.0011946
Random Forest No GenHydro	0.0024508

In this last method, RMSE is improved by reducing the model and using only the most significant variables. The 4th method introduced in regression trees without GenHydro does not improve the model at all.

Results

The following table shows the RMSE's obtained in the training of the different models.

Model	RMSE
Linear Regression All Variables	0.0005619
Linear Regression 9 Variables	0.0015678
Linear Regression 6 Variables	0.0021127
Regression Tree All Variabless	0.0674774
Regression Tree 9 Variables	0.0674774
Regression Tree 6 Variables	0.0674774
Regression Tree No GenHydro	0.1590990
Random Forest All Variables	0.0232776
Random Forest 9 Variables	0.0011295
Random Forest 6 Variables	0.0011946
Random Forest No GenHydro	0.0024508

As it was mentioned before, each technique was affected by the uniqueness of its model. Linear regression RMSE improves with more variables, but does not improve its predictive power. Regression trees are clouded by a very strong variable such as GenHydro that clouds all other variables from showing on the trees, yet removing the variable increases the RMSE. In the case of random forests,averaging several regression trees helps to not be clouded by GenHydro, but also removing it from the model increases RMSE.

Finally, we can see that 2 variables that remain strong and constant all along the excercise are CumlFinancial and GenTotal, or Financial Incentives and total energy generation. This are good matches that can be generalized for real life applications as countries that are big energy producers have a better chance to be renewable energy producers and even more important, countries that support energy generation financially set strong basis for entrepreneurs to launch energy generation projects.

Conclusion

On this project: There were many trade offs while carrying out this project. The first was the choice of data set, where personal interest clashed quality of the data. As a first choice I had chosen crime, but even though there were enough observations, the variables such as neighborhood, month, day, hour, minute, latitude, longitude and type of crime there was not much that could be predicted. This was a great data set for exploration and visualization but without much predictive power. As a second choice, a data set on real estate prices in New York City was the opposite, big potential for predictions, but not much interest on my side for this topic. Finally settling for a data set on a topic in which I am deeply interested, yet did not contain complete technical information on the variables to make quantitative assertions.

On this data set: The choice of dependent variable was another trade off, I wanted to see the predictors of renewable energy generation, yet the variable being continuous, only a few machine learning techniques could be applied (linear regression, regression trees and random forest). Had I chosen a different variable (categorical), I could have ran logistic regression, LDA, LQA or KNN techniques. On the results obtained, we can see how each machine learning technique behaves. Linear regression has a lower RMSE the more variables we include in the model, which does not make it necessarily more predictive. Regression trees can be clouded by one variable (GenHydro) and avoid visibility, yet removing the variable hurts the predictive power of the model. Finally, random forests profit from the strength of combining many trees to avoid being clouded by one variable, yet removing such variable also hurts the model.

Not having technical information on the variables allowed to make only numerical assumptions on relationships between the variables. It was still very insightful on the relationship of energy predictors according to theory and the results shed by this data set. Theory says that government intervention will forcefully change the behavior of the private sector to produce energy through government incentives, this will in turn provoke a decrease in prices due to the increased supply of energy, and finally cause a surge in consumption since energy becomes cheaper. This data set shows that some of these relationships are not as straightforward, Government Incentives and Energy Generation Government incentives slightly increased in the year 2000s, seeing a surge in 2005 and an even bigger push in 2009. An increase in renewable energy generation did happen, but the reason is not in this data set ((personally I think it was wind energy), because solar remained low (it started growing around 2014, time after the scope of this data set, due to a drop in solar technology prices), and hydroelectric energy had ups and downs but ended at lower levels in 2013 than in 2000. Energy Generation and Prices More energy supply did not reduced prices. It didn't in the past and it probably wont do in the near future. It didn't in the past because the initial investment of launching new energy projects is very high and projects need to keep charging high prices to recover the investment. It will not in the future because lowering energy prices too much but cause a double problem, on one side increased consumption on times where efforts are made to reduce consumption, and lower return on investment for new projects wishing to launch. Prices and consumption Regardless of increased prices, energy consumption continues to rise, and it will continue to do so and we become more energy dependent for most of human daily activities. The introduction of electric mobility is one of the big drivers of this increase and it is not necessarily negative as long as the energy generated to power this vehicles is renewable energy. The decrease in industrial energy can be explained because we are using sales as a proxy for consumption and in this case, there was being an effort for many industries (data centers, tech, car manufacturers) to create their own energy, therefore stop "buying" it from utilities. Lastly, the final question of income causing more energy consumption, we can positively see that income has risen a lot more than energy consumption, which shows that even though people are earning more, they are being careful with their consumption, which is a reflection of energy consciousness.

References

Irizarry, Rafael A., Introduction to Data Science, Data Analysis and Prediction Algorithms with R, 2021-07-03, https://rafalab.github.io/dsbook/

https://energypost.eu/green-new-deal-can-learn-from-obamas-90bn-clean-energy-plan-of-2009/